Flight Delay Prediction Using Machine Learning Algorithm

Hemadri A D[1]. Kumar Raja D R[2]

[1,2] School of Computing and Information technology, REVA University, India.

Received: 12 July 2022 / Revised: 02 Aug 2022 / Accepted: 28 Aug 2022

Abstract – The current and the existing circumstances due to the traffic congestion causing flight delays these flight delays not only causing economic impact but also have harmful environment effects and degrading the passenger quality of service and fuel consumption and gas consumption the airline management had becoming the increasingly challenging to overcome this issues. By using the factors causing the airline delay we carry out the predictive analysis and machine learning algorithms to find the causes of flight delays.

Index Terms – Airport congestion, Algorithm, Dataset, Impact on passenger, Machine learning, Predict the causes, Risk analysis.

I. INTRODUCTION

Flights are taken a dominant part in the current day to day life in travelling many of the travelers prefer flights for travelling due to their speed,time management and comfort .This had led a phenomenal growth in the usage of flights due to the demand on travelling in flights .A flight delay is said to occur when a air line lands or take off later then departure arrival time or landing time when the arrival time is greater than the given time this is considered as a flight delay Notable delays are due to weather conditions, traffic condition ,late reaching of flight due to previous flight,mantainance and security issues

These delays loses its trust on such famous recognized airlines. In the paper we have proposed a model based on machine learning algorithm to predict the delays of flight by the data set provided by the department of airlines that offers information of delays of flight. we have used polynomial regression and linear regression to predict the airline delays.

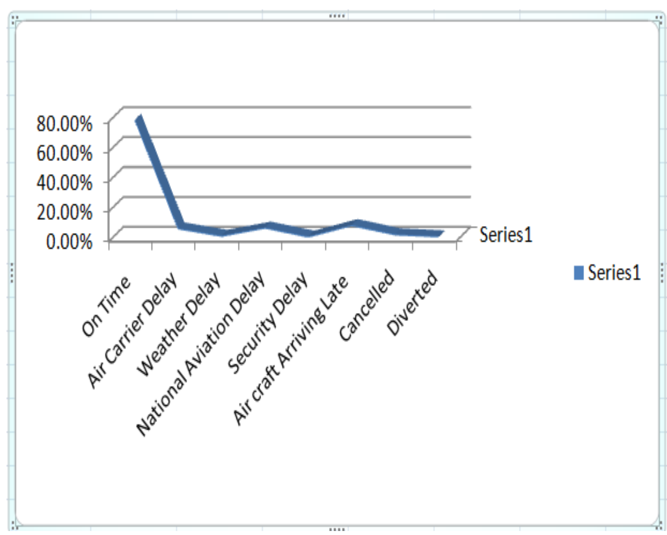

Fig 1 Delay Statistics of a year

II. PROPOSED MODEL

The proposed system The data base consist of the Airline-Out Time ,Schedule Time ,Airline-ln Time and Air borne Time collected from the domestic flights by the large airline carriers the summary air line contains on time flight, delayed flight, canceled flight, and the diverted flight

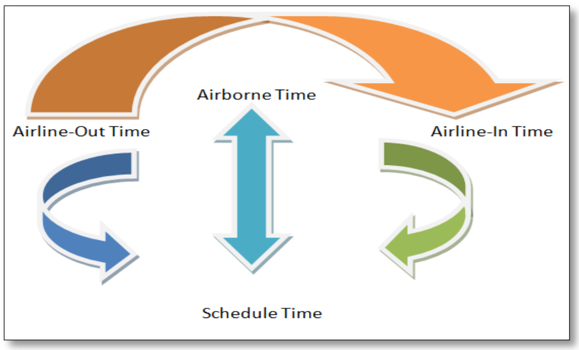

Fig 2. Proposed model

These reports data is used to predict which air lines are better to travel to avoid significant delays.
Advantages
1. The Model helps to know which flight is better to fly.
2. The model gives the information of all the delayed flights.

III. METHOLOGY

The data was obtained from a reputable online government department database that offers information on air traffic delays in the united states. The Bureau of Transportation statistics (BTS) of the United states Department of Transportation (DOT).

The test samples Obtained from the Bureau of Transportation statics (BTS)and United states department of transportation (DOT) generates the test samples used to generate the report of the reasons of flight delays the collected data contains ,Year,Month,Day,Day-of-week, Airline ,Flight-Number,Tail-Number,Origin-Airport,Destination-Airport,Scheduled-Departure,Departure-Time,Departure-Delay,Taxi-out,Wheels Off,Scheduled-time,Elapsed-time,Air-time,Distance,Wheels-on,Taxi-In,Scheduled-Arrival,Arrival-Time,Arrival-Delay,Diverted,Cancelled,Cancelled-Reason,Air-system-Delay,Taxi-Out,Wheels off .

These data set obtained from transportation are Normalized and Filtered and undergo cleaning. The obtained data is used for prediction of flight delays.

ATTRIBUTE	DESCRIPTIONS OF ATTRIBUTES
YEAR,MONTH,DAY, DAY_OF _ WEEK	Dates of the flight
AIRLINES	It is the IATA Code in Identify Unique airlines
ORIGIN_AIRPORT and DESTINATION AIRPORT	Code attributed by IATA to identity the airport
SCHEDULED_DEPARTURE and SCHEDULED_ARRIVAL	Scheduled times of take-off and landing
DEPARTMENT_TIME and AARIVAL_TIME	Real times at which take-off and landing took place
DEPARTURE_DELAY and ARRIVAL_DELAY	Difference in minutes between planned and real times distance in miles

Fig 3 Attribute of data set

	year	Month	Day	Day-of-week	Airline	Flight-Number	Tail-Number	Origin-Airport	Destination-Airport	Scheduled-Departure	Departure-Time	Departure-Delay	Taxi-OUT	Wheels off
Column type	Int64	Int64	Int64	Int64	object	Int64	object	object	Object	Int64	Float64	Float 64	Float64	Float64
Null values (nb)	0	0	0	0	0	0	14721	0	0	0	86153	86153	89047	89047
Null values %	0	0	0	0	0	0	0.252978	0	0	0	1.48053	1.48053	1.53026	1.53036

Fig 4 Data set 01

	Scheduled-Time	Elapsed-time	Air-time	Distance	Wheels-on	Taxi-IN	Scheduled-Arrival	Arrival-Time	Arrival-Delay	Diverted	cancelled	Cancelled-reason	Air-system-delay	Security-delay
Column type	Float64	Float64	Float64	Int64	Float64	Int64	Int64	Float64	Float64	Int64	Int64	object	Float64	Float64
Null values (nb)	6	105071	105071	0	92513	92513	0	92513	105071	0	0	5729195	4755640	4755640
Null values %	0.000103109	1.80563	1.80563	0	1.58982	1.58982	0	1.58982	1.80563	0	0	98.4554	81.725	81.725

Fig 5 Data Set 02

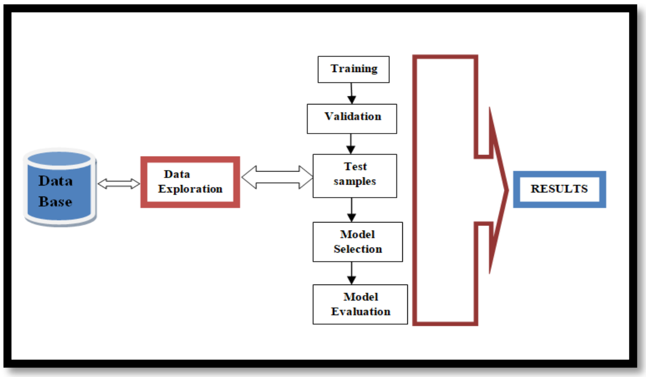

Fig. 6 Methodology

1. **Data Base**
   The Airline authority collects the data of Air line –in, Airline–out, delayed flight, diverted flight for the security purpose the data set is collected from the airline authority for prediction of flight delays .
2. **Data Exploration**
   This process involves collection of data set new attribute. This data exploration contains the causes of flight delays .i.e canceled flight, weather delays.
3. **Training**
   Training data involves cleaning and normalization of the attributes. Due to the nature of attributes of air traffic, most of the flights are not delayed these data reported in the data set are skewed and they are not equally represented. There fore to prevent the baised data set we use this model of training.
4. **Validation**
   The trained data set is splitted into to two data set one is for test sample and the validation is done for other sample to prevent the model from overfitting .
5. **Test samples**
   The test samples are collected from one set of the validation data set the test data set is unbiased model for evaluation of a final model fit .
6. **Model Selection**
   When your data has different values and even different conditions it is difficult to predict the results using the data set. In the paper we have used two machine learning algorithm

1.Polynomial regression
Polynomial regression data points clearly fits the data points the relation ship with X and Y it finds the best way to draw a line through the data points.

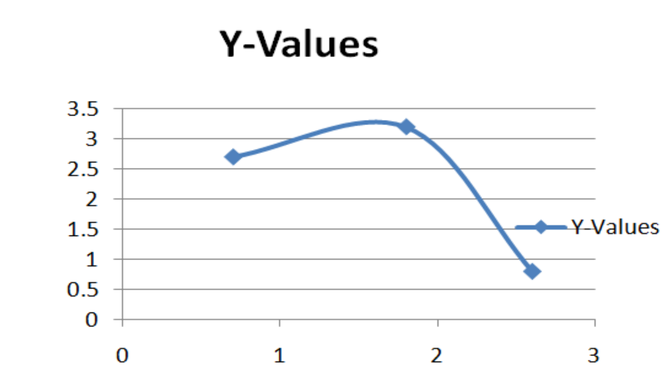

Fig. 7 Polynomial regression

2.Linear regression.
Linear regression algorithm can find the results for more than one independent values.we can predict the values of more than two or three attributes.

Flight Delay	Flight Name	Flight Arrival	Flight Take off	Delay Flight	Date	Time

Fig 8 Linear regression

These algorithm is used to find a relationship between the two data points.

IV. RESULTS

	DAY_OF_MONTH	DAY_OF_WEEK	OP_CARRIER_AIRLINE_ID	ORIGIN_AIRPORT_ID	DEST_AIRPORT_ID	DEP_TIME
0	1	2	20363	11953	10397	601.0
1	1	2	20363	13487	11193	1359.0
2	1	2	20363	11433	11193	1215.0
3	1	2	20363	15249	10397	1521.0
4	1	2	20363	10397	11778	1847.0

Fig. 9 Result of data set

DEP_TIME	DEP_DEL15	ARR_TIME	DIVERTED	DISTANCE
601.0	0.0	722.0	0.0	300.0
1359.0	0.0	1633.0	0.0	596.0
1215.0	0.0	1329.0	0.0	229.0
1521.0	0.0	1625.0	0.0	223.0
1847.0	0.0	1940.0	0.0	579.0

Fig. 10 Result of data set

```
 #   Column                 Non-Null Count   Dtype
---  ------                 --------------   -----
 0   DAY_OF_MONTH           583985 non-null  int64
 1   DAY_OF_WEEK            583985 non-null  int64
 2   OP_UNIQUE_CARRIER      583985 non-null  object
 3   OP_CARRIER_AIRLINE_ID  583985 non-null  int64
 4   OP_CARRIER             583985 non-null  object
 5   TAIL_NUM               581442 non-null  object
 6   OP_CARRIER_FL_NUM      583985 non-null  int64
 7   ORIGIN_AIRPORT_ID      583985 non-null  int64
 8   ORIGIN_AIRPORT_SEQ_ID  583985 non-null  int64
 9   ORIGIN                 583985 non-null  object
 10  DEST_AIRPORT_ID        583985 non-null  int64
 11  DEST_AIRPORT_SEQ_ID    583985 non-null  int64
 12  DEST                   583985 non-null  object
 13  DEP_TIME               567633 non-null  float64
 14  DEP_DEL15              567630 non-null  float64
 15  DEP_TIME_BLK           583985 non-null  object
 16  ARR_TIME               566924 non-null  float64
 17  ARR_DEL15              565963 non-null  float64
 18  CANCELLED              583985 non-null  float64
 19  DIVERTED               583985 non-null  float64
 20  DISTANCE               583985 non-null  float64
dtypes: float64(7), int64(8), object(6)
```

Fig. 11 Loading of data set

```
LinregressResult(slope=0.738785248054155, intercept=0.05759336951361887 5, rvalue=0.71942
99212037087, pvalue=0.0, stderr=0.0009480896010342847)
```

Fig. 12 Result of Linear regression

```
LinregressResult(slope=2.2374154184666654e-06, intercept=0.18412145596027452, rvalue=0.0
034065404631858473, pvalue=0.010384391665888941, stderr=8.730463522779241e-07)
```

Fig. 13 Result of Linear regression

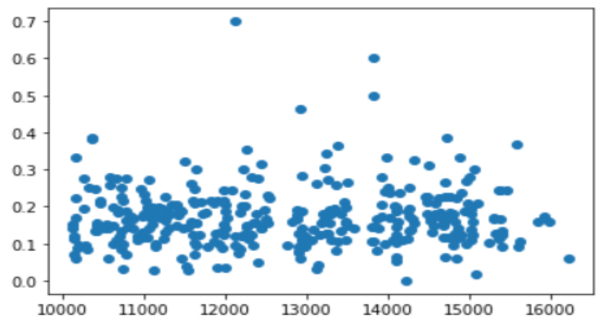

Fig. 14 Result of polynomial regression

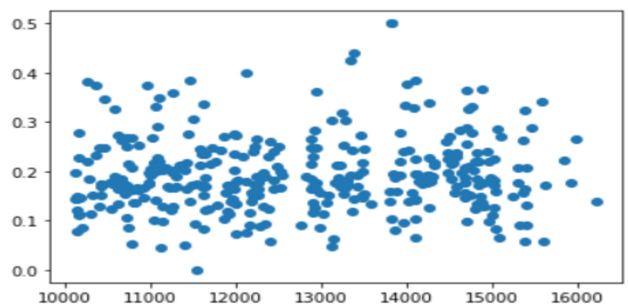

Fig. 15 Result of polynomial regression

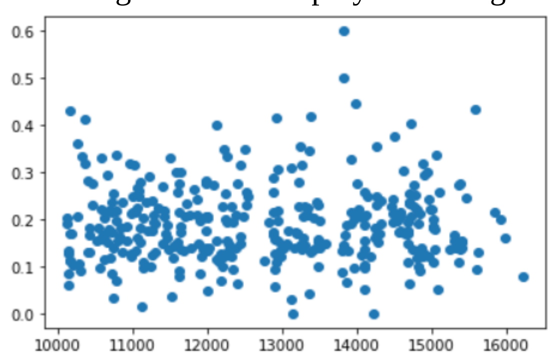

Fig. 16 Result of polynomial regression

V. CONCLUSION

In the existing paper the flight delay prediction has few or other draw back in their algorithms, data model. In this paper we have overcome those draw back and provided the accurate result and prediction using linear and polynomial regression.

REFERENCE

1. Sternberg, A., Soares, J., Carvalho, D., & Ogasawara, E. (2017). A review on flight delay prediction. *arXiv preprint arXiv:1703.06118*.
2. Borse, Y., Jain, D., Sharma, S., Vora, V., & Zaveri, A. (2020). Flight Delay Prediction System. *Int. J. Eng. Res. Technol*, *9*(3), 88-92.
3. Gui, G., Liu, F., Sun, J., Yang, J., Zhou, Z., & Zhao, D. (2019). Flight delay prediction based on aviation big data and machine learning. *IEEE Transactions on Vehicular Technology*, *69*(1), 140-150.
4. Kim, Y. J., Choi, S., Briceno, S., & Mavris, D. (2016, September). A deep learning approach to flight delay prediction. In *2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC)* (pp. 1-6). IEEE.
5. Yazdi, M. F., Kamel, S. R., Chabok, S. J. M., & Kheirabadi, M. (2020). Flight delay prediction based on deep learning and Levenberg-Marquart algorithm. *Journal of Big Data*, *7*(1), 1-28.
6. Yu, B., Guo, Z., Asian, S., Wang, H., & Chen, G. (2019). Flight delay prediction for commercial air transport: A deep learning approach. *Transportation Research Part E: Logistics and Transportation Review*, *125*, 203-221.
7. Cai, K., Li, Y., Fang, Y. P., & Zhu, Y. (2021). A deep learning approach for flight delay prediction through time-evolving graphs. *IEEE Transactions on Intelligent Transportation Systems*.
8. Zhu, X., & Li, L. (2021). Flight time prediction for fuel loading decisions with a deep learning approach. *Transportation Research Part C: Emerging Technologies*, *128*, 103179.
9. Sreedhar Kumar, S., Ahmed, S. T., & NishaBhai, V. B. Type of Supervised Text Classification System for Unstructured Text Comments using Probability Theory Technique. *International Journal of Recent Technology and Engineering (IJRTE)*, *8*(10).
10. Parveen, A., Ahmed, S. T., Gulmeher, R., & Fatima, R. (2021). VANET’s Security, Privacy and Authenticity: A Study.